Survey on Emotional Body Gesture Recognition

Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari,

Abstract—Automatic emotion recognition has become a trending research topic in the past decade. While works based on facial expressions or speech abound recognizing affect from body gestures remains a less explored topic. We present a new comprehensive survey hoping to boost research in the field. We first introduce emotional body gestures as a component of what is commonly known as "body language" and comment general aspects as gender differences and culture dependence. We then define a complete framework for automatic emotional body gesture recognition. We introduce person detection and comment static and dynamic body pose estimation methods both in RGB and 3D. We then comment the recent literature related to representation learning and emotion recognition from images of emotionally expressive gestures. We also discuss multi-modal approaches that combine speech or face with body gestures for improved emotion recognition. While pre-processing methodologies (e.g. human detection and pose estimation) are nowadays mature technologies fully developed for robust large scale analysis, we show that for emotion recognition the quantity of labelled data is scarce, there is no agreement on clearly defined output spaces and the representations are shallow and largely based on naive geometrical representations.

Index Terms—emotional body language, emotional body gesture, emotion recognition, body pose estimation, affective computing

◆

1 INTRODUCTION

DURING conversations people are constantly changing nonverbal clues, communicated through body movement and facial expressions. The difference between the words people pronounce and our understanding of their content comes from nonverbal communication also commonly called body language. Some examples of body gestures and postures, key components of body language are shown in Fig. 1.

Although it is a significant aspect of human social psychology, the first modern studies concerning body language has become popular in 1960s [1]. Probably the most important work published before 20th century was *The Expression of the Emotions in Man and Animals* by Charles Darwin [2]. This work is the foundation of modern approach to body language and many of Darwin's observations were confirmed by subsequent studies. Darwin observed that people all over the world use facial expressions in a fairly similar manner. Following this observation, Paul Ekman researched patterns of facial behavior among different cultures of the world. In 1978, Ekman and Friesen developed the Facial Action Coding System (FACS) to model human facial expressions [3]. In an updated form, this descriptive

Figure 1. Body language includes different types of nonverbal indicators such as facial expressions, body posture, gestures and eye movements. These are important markers of the emotional and cognitive inner state of a person. In this work, we review the literature on automatic recognition of body expressions of emotion, a subset of body language that focuses on gestures and posture of the human body. The images have been taken from [4].

anatomical model is still being used in emotion expressions recognition.

The study of use of body language for emotion recognition was conducted by Ray Birdwhistell who found that the final message of an utterance is affected only 35% by the actual words and 65% by non-verbal signals [5]. In the same work, analysis of thousands of negotiations recordings revealed that the body language decides the outcome of those negotiations in 60% - 80% of cases. The research also showed that during a phone negotiation, stronger arguments win, however during a personal meeting, decisions are made on the basis of what we see rather than what we hear [1]. At the

- F. Noroozi and G. Anbarjafari are with the iCV Research Group, Institute of Technology, University of Tartu, Tartu, Estonia.
  E-mail: {fatemeh.noroozi,shb}@ut.ee
- D. Kamińska and T. Sapiński are with Department of Mechatronics, Lodz University of Technology, Lodz, Poland.
  E-mail: dorota.kaminska@p.lodz.pl, sapinski.tomasz@gmail.com
- C. A. Corneanu and S. Escalera are with the University of Barcelona and Computer Vision Center, Barcelona, Spain.
  E-mail: cipriancorneanu@gmail.com, sergio@maia.ub.es
- G. Anbarjafari is also with Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gaziantep, Turkey.

present, most researchers agree that words serve primarily to convey information and the body movements to form relationships and sometimes even to substitute the verbal communication (e.g. lethal look).

Gestures are one of the most important forms of non-verbal communication. They include movements of hands, head and other parts of the body that allow individuals to communicate a variety of feelings, thoughts and emotions. Most of the basic gestures are the same all over the world: when we are happy we smile when we are upset we frown [6], [7], [8].

According to [1], gestures can be of the following types:

- Intrinsic: For example, nodding as a sign of affir-mation or consent is probably innate, because even people who are blind from birth use it;
- Extrinsic: For example, turning to the sides as a sign of refusal is a gesture we learn during early childhood - It happens when, for example, a baby has had enough milk from the mother's breast, or with older children when they refuse a spoon during feeding;
- A result of natural selection: For example, the ex-pansion of the nostrils to oxygenate the body can be mentioned, which takes place when preparing for battle or escape.

The ability to recognize the attitude and thoughts from one's behavior was the original system of communication before the speech. Understanding of emotional state enhances the interaction. Although computers are now a part of human life, the relation between a human and a machine is not natural. Knowledge of the emotional state of the user would allow the machine to adapt better and generally improve cooperation.

While emotions can be expressed in different ways, auto-matic recognition has mainly focused on facial expressions and speech. About 95% of the literature dedicated to this topic focused on faces as a source for emotion analysis [9]. Considerably less works were done on body gestures and posture. With recent developments of motion capture technologies and reliability, the literature about automatic recognition of expressive movements grew significantly.

Despite the increasing interest in this topic, we are aware of just a few relevant survey papers. For example, Klein-smith et al. [10] reviewed the literature on affective body expression perception and recognition with an emphasis on inter-individual differences, impact of culture and multi-modal recognition. In another paper, Kara et al. [11] intro-duced categorization of movement into four types: commu-nicative, functional, artistic, and abstract and discussed the literature associated with these types of movements.

In this work, we cover all the recent advancements in au-tomatic emotion recognition from body gestures. The reader interested in emotion recognition from facial expressions or speech is encouraged to consult dedicated surveys [12], [13], [14]. In this work we refer to these only marginally and only as complements to emotional body gestures. In Sec. 2 we briefly introduce key aspects of affect expression through body language in general and we discuss in-depth cultural and gender dependency. Then, we define a standard pipeline for automatically recognizing body gestures of

emotion in Sec. 4 and we discuss in details technical aspects of each component of such pipeline. Furthermore, in Sec. 5 we provide a comprehensive review of publicly available databases for training such automatic recognition systems. We conclude in Sec. 6 with discussions and potential future lines of research.

2 EXPRESSING EMOTION THROUGH BODY LAN-GUAGE

According to [15], [16] body language includes different kinds of nonverbal indicators such as facial expressions, body posture, gestures, eye movement, touch and the use of personal space. The inner state of a person is expressed through elements such as iris extension, gaze direction, posi-tion of hands and legs, the style of sitting, walking, standing or lying, body posture, and movement. Some examples are presented in Fig. 1.

After the face, hands are probably the richest source of body language information [17], [18]. For example, based on the position of hands one is able to determine whether a person is honest (one will turn the hands inside towards the interlocutor) or insincere (hiding hands behind the back). Exercising open-handed gestures during conversation can give the impression of a more reliable person. It is a trick often used in debates and political discussions. It is proven that people using open-handed gestures are perceived posi-tively [1].

Head positioning also reveals a lot of information about emotional state. The research [6] indicates that people are prone to talk more if the listener encourages them by nodding. The pace of the nodding can signal patience or lack of it. In neutral position the head remains still in front of the interlocutor. If the chin is lifted it may mean that the person is displaying superiority or even arrogance. Exposing the neck might be a signal of submission. In [2] Karl Darwin noted that like animals, people tilt their heads when they are interested in something. That is why women perform this gesture when they are interested in men, an additional display of submission results in greater interest from the opposite sex, e.g. a lowered chin signals a negative or aggressive attitude.

The torso is probably the least communicative part of the body. However, its angle with the body is an indicative attitude. For example placing the torso frontally to the interlocutor can be considered as a display of aggression. By turning it at a slight angle one may be considered self-confident and devoid of aggression. Leaning forward, especially when combined with nodding and smiling, is the most distinct way to show curiosity [6].

The above considerations indicate that in order to cor-rectly interpret body language as indicators of emotional state, various parts of body must be considered at the same time. According to [19], body language recognition systems may benefit from a variety of psychological behavioral protocols. An example of general movements protocol for six basic emotions is presented in Table 1.

2.1 Culture differences

It has been reported that gestures are strongly culture-dependent [23], [24]. However, due to exposure to mass-

Table 1
The general movement protocols for the six basic emotions [20], [21], [22].

Emotion	Associated body language
Fear	Noticeably high heart beat-rate (visible on the neck). Legs and arms crossing and moving. Muscle tension: Hands or arms clenched, elbows dragged inward, bouncy movements, legs wrapped around objects. Breath held. Conservative body posture. Hyper-arousal body language.
Anger	Body spread. Hands on hips or waist. Closed hands or clenched fists. Palm-down posture. Lift the right or left hand up. Finger point with right or left hand. Finger or hand shaky. Arms crossing.
Sadness	Body dropped. Shrunk body. Bowed shoulders. Body shifted. Trunk leaning forward. The face covered with two hands. Self-touch (disbelief), body parts covered or arms around the body or shoulders. Body extended and hands over the head. Hands kept lower than their normal positions, hands closed or moving slowly. Two hands touching the head and moving slowly. One hand touching the neck. Hands closed together. Head bent.
Surprise	Abrupt backward movement. One hand or both of them moving toward the head. Moving one hand up. Both of the hands touching the head. One of the hands or both touching the face or mouth. Both of the hands over the head. One hand touching the face. Self-touch or both of the hands covering the cheeks or mouth. Head shaking. Body shift or backing.
Happiness	Arms open. Arms move. Legs open. Legs parallel. Legs may be stretched apart. Feet pointing something or someone of interest. Looking around. Eye contact relaxed and lengthened.
Disgust	Backing. Hands covering the neck. One hand on the mouth. One hand up. Hands close to the body. Body shifted. Orientation changed or moving to a side. Hands covering the head.

media, there is a tendency of globalization of some gestures especially in younger generations [6]. This is despite the fact that the same postures might have been used for expressing significantly different feelings by their previous generations. Consequently, over time, some body postures might change in meaning, or even disappear. For instance, the thumb-up symbol might have different meanings in different cultures. In Europe it stands for number "1" in Japan for "5", while in Australia and Greece, using it may be considered insulting. However, nowadays, it is widely used as a sign of agreement, consent or interest [25].

Facial expressions of emotion are similar across many cultures [26]. This might hold in the case of postures as well. In [27], the effect of culture and media on emotional expressions was studied. One of the conclusions was that an American and a Japanese infant present closely similar emotional expressions. Most of the studies reported on this topic in the literature inferred that intrinsic body language, gestures and postures are visibly similar throughout the world. However, a decisive conclusion still requires more in-depth exploration, which is challenging due to the variety of topics that need to be studied on numerous cultures and countries. Therefore, the researchers investigating this issue prefer to concentrate on a certain activity, and study it on various cultures, which may lead to a more understandable distinction. For example, in many cultures, holding hands resembles mutual respect, but in some others touching one another in exchanging greetings might not be considered usual [25].

2.2 Gender differences

Women are believed to be more perceptive than men due to the concept of female intuition [29]. There are some fundamental differences in the way women and man communicate through body language [28] (see Fig. 2 for some trivial examples). This may be caused by influence of culture (tasks and expectations that face both sexes), body composition, makeup and worn type of clothes.

Women wearing mini skirts often sit with crossed legs or ankles. But it is not the sole reason of this gesture applying

Figure 2. There are fundamental differences in the way men and women communicate through body language. In certain situations, one can easily discriminate the gender when only the body pose is shown. Illustration from [28].

almost exclusively to women. As a result of body composition, most men are not able to sit that way, thus this position became a symbol of femininity. Another good example is the cowboy pose popularized by old western movies. In this pose the thumbs are placed in the belt loops or pockets with the remaining fingers pointed downwards towards the crotch. Men use this gesture when they defend their territory or while demonstrating their courage. A similar position has been also observed among monkeys [1].

Generally, women show emotions and their feelings more willingly than men [30], which are associated with qualities such as kindness, supportiveness, affection and care for others. Men are more likely to display power and dominance while simultaneously hiding the melting mood [30]. However, nowadays these general tendencies start to faint and are considered as gender stereotypes [31].

3 MODELS OF THE HUMAN BODY AND EMOTION

Before discussing the main steps for automatically recognizing emotion from body gestures, (details in Sec. 4), we first discuss modelling of the input and output of such systems. The input will be an abstraction of the human body (and
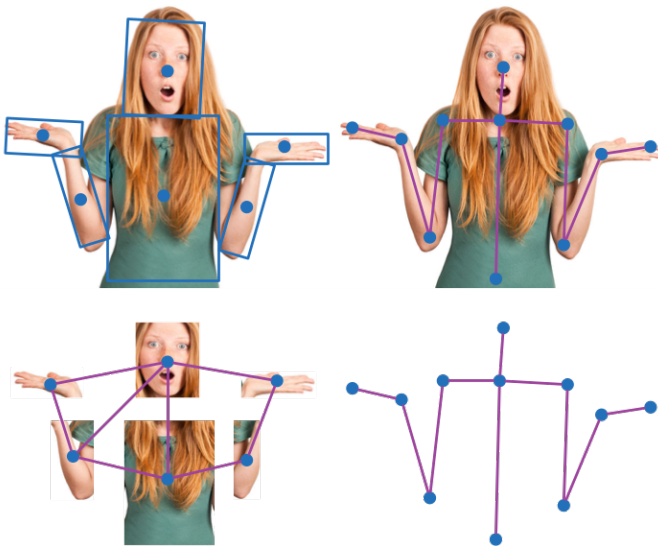
Figure 3. The two most common ways of modelling the human body in automatic processing are either as an ensemble of body parts or as a kinematic model. In ensemble of body parts (left) different parts of the body are independently detected and soft restrictions can be imposed to refine these detections. A kinematic model (right) is a collections of interconnected joints with predefined degrees of freedom similar to the human skeleton.

Figure 4. Example of body pose estimation and tracking using ensemble of parts namely, head and hands [35].

its dynamics) that we would like to map through machine learning methods to a certain predefined abstraction of emotion. Deciding the appropriate way of modelling the human body and emotion is an essential design decision. We begin by discussing abstractions of the human body (Sec. 3.1) and then main models of emotion used in affective computing (Sec. 3.2).

3.1 Models of the human body

Human body has evolved such that it can perform complex actions, which require coordination of various organs. Therefore, many everyday actions present unique spatio-temporal movement structures [32]. In addition, some pairs of body actions, e.g. walking and drinking, may be performed at the same time and their expression might not be additive [33]. The two main lines for abstracting the human body have been following either a constrained composition of human body parts or a kinematic logic based on the skeletal structure of the human body (see Fig. 3).

Part Based Models. In a part based approach the human body is represented as flexible configuration of body parts. Body parts can be detected independently (face, hands, torso) and priors can be imposed using domain knowledge of the human body structure to refine such detection. Some examples of ensemble of parts models of the human body are pictorial structures and grammar models. Pictorial structures are generative 2D assemblies of parts, where each part is detected with its specific detector. Pictorial structures are a general framework for object detection widely used for people detection and human pose estimation [34]. An example of body pose estimation using pictorial structures is shown in Fig. 4.

Grammar models provide a flexible framework for detecting objects, which was also applied for human detection
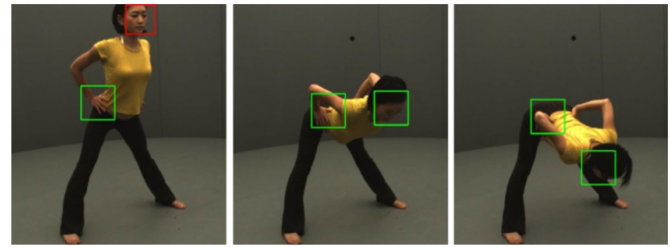
in [36]. Compositional rules are used to represent objects as a combination of other objects. In this way, human body could be represented as a composition of trunk, limbs and face; as well composed by eyes, nose and mouth.

Kinematic Models. Another way of modelling the human body is by defining a collection of interconnected joints also known as kinematic chain models. This is usually a simplification of the human skeleton and its mechanics. A common mathematical representation of such models is through a cyclical tree graphs which also present the advantage of being computationally convenient. Contrary to part based approach [34], nodes of structure trees represent joints, each one parameterized with its degrees of freedom. Kinematic models can be planar, in which case they are a projection in the image plane or depth information can be considered as well. Richer, more realistic variants can be defined for example as a collection of connected cylinders or spheroids or 3D meshes. Examples of body pose detection using kinematic models and deep learning methods are shown in Fig. 5.

3.2 Models of emotion

The best way of modelling affect has been subject of debate for a long time and many perspectives upon the topic were proposed. The most influential models (and in general most relevant for affective computing applications) can be classified in three main categories: categorical, dimensional and componential [42] (see Fig. 6 for examples of each category).

Categorical models. Classifying emotions into a set of distinct classes that can be recognized and described easily in daily language has been common since at least the time of Darwin. More recently, influenced by the research of Paul Ekman [43], [46] a dominant view upon affect is based on the underlying assumption that humans universally express and recognize a set of discrete primary emotions which include happiness, sadness, fear, anger, disgust, and surprise. Mainly because of its simplicity and its universality claim, the universal primary emotions hypothesis has been by far the first choice for affective computing research and has been extensively exploited.

Dimensional models. Another popular approach is to model emotions along a set of latent dimensions [44], [47], [48]. These dimensions include valence (how pleasant or unpleasant a feeling is) activation (how likely is the person to take action under the emotional state) and control (the sense of control over the emotion). Due to their continuous nature, such models can theoretically describe more complex and subtle emotions. Unfortunately, the richness of the space
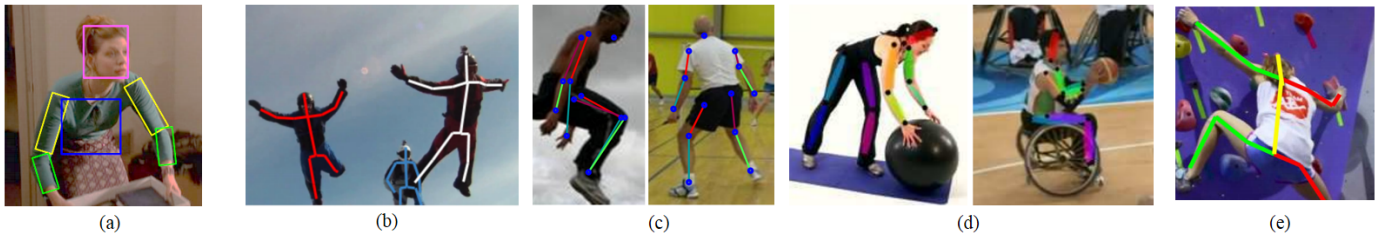
Figure 5. Examples of body pose estimation using artificial neural networks: (a) latent structured support vector machines LSSVM [37], (b) Associative embedding supervised convolutional neural network (CNN) for group detection [38], (c) Hybrid architecture consisting of a deep CNN and a Markov Random Field [39], (d) Replenishing back-propagated gradients and conditioning the CNN procedure [40] and (e) CNN by using the iterative error feedback processing [41].
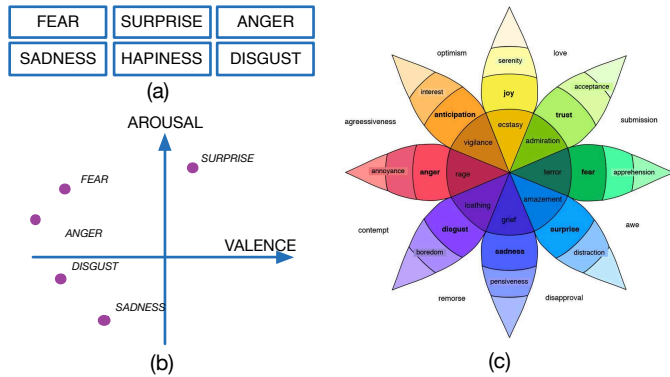

Figure 6. There are three main ways of modelling emotion in affective computing. Emotion can be defined (a) in a categorical way (here a universal set of emotions as define by Ekman [43]) (b) as point in space defined along a set of latent dimensions (Russells model depicted) [44] or (c) in a hybrid manner (Plutchik model shown) [45].

4 BODY GESTURE BASED EMOTION RECOGNITION

In this section, we present the main components of what we call an Emotion Body Gesture Recognition (EBGR) system. For a detailed depiction see Fig. 7. An important preparation step, which influences all the subsequent design decisions for such an automatic pipeline is the determination of the appropriate modelling of input (human body) and targets (emotion). Depending on the type of the model that has been chosen, either a publicly accessible database can be utilized, or a new one needs to be created. Similarly, other elements of the system need to be selected and configured such that they are compatible with each other, and overall, provide an efficient performance. Regardless of the foregoing differences between various types of EBGR systems, the common first step is to detect the body as a whole, i.e. to subtract the background from every frame which represents a human presenting a gesture. We will briefly discuss the main literature for human detection in Sec. 4.1. The second step is detection and tracking of the human pose in order to reduce irrelevant variation of data caused by posture (we dedicate Sec. 4.2 to this). The final part of the pipeline, which we discuss in Sec. 4.3, consists in building an appropriate representation of the data and applying a learning technique (usually classification or regression) to map this representation to the targets. We conclude this section with a presentation of the most important applications of automatic recognition of emotion using body gesture.

4.1 Human Detection

Human detection in images usually consists in determining rectangular bounding boxes that enclose humans. It can be a challenging task because of the non-rigid nature of the human body, pose and clothing, which result in high variation of appearance. In uncontrolled environments changes of illumination and occlusions add to the complexity of the problem.

A human detection pipeline follows the general pipeline of object detection problems and consists of extracting potential candidate regions, representing those regions, classifying the regions as human or non-human, and merging positives into final decisions [50]. If depth information is available, it can be used to limit the search space and considerably simplify the background substraction problem [50]. Modern techniques might not exactly follow this modularization, either by jointly learning representation

is more difficult to use for automatic recognition systems because it can be challenging to link such described emotion to a body expression of affect. This is why, many automatic systems based on dimensional representation of emotion simplified the problem by dividing the space in a limited set of categories like positive vs negative or quadrants of the 2D space [49].

Componential models. Somehow in-between categorical and dimensional models in terms of descriptive capacity, componential models of affect, arrange emotions in a hierarchical fashion where each superior layer contains more complex emotions which can be composed of emotions of previous layers . The best example of componential models was proposed by Plutchik [45]. According to his theory, more complex emotions are combinations of pairs of more basic emotions, called dyads. For example, love is considered to be a combination of joy and trust. Primary dyads, e.g. optimism=anticipation+joy, are often felt, secondary dyads, e.g. guilt=joy+fear, are sometimes felt and tertiary dyads, e.g. delight=joy+surprise, are seldom felt. These types of models are rarely used in affective computing literature compared to the previous two but should be taken into consideration due to their effective compromise between ease of interpretation and expressive capacity.
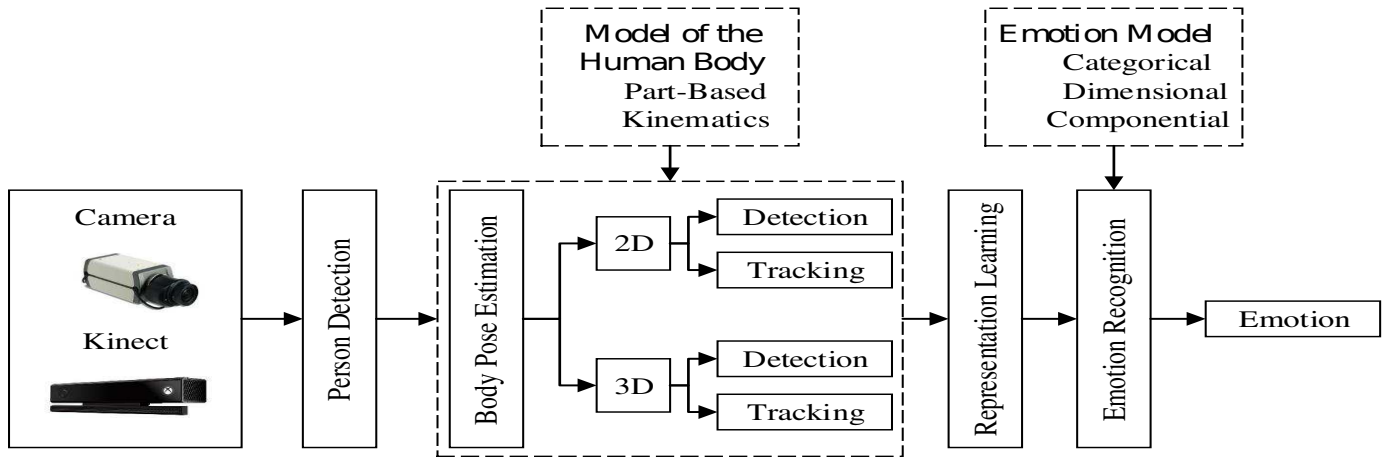
Figure 7. General overview of an Emotion Body Gesture Recognition system. After detecting persons in the input for background extraction, a common step is to estimate the body pose. This is done either by detecting and tracking different parts of the body (hands, head, torso, etc.) or by mapping a kinematic model (a skeleton) to the image. Based on the extracted model of the human body, a relevant representation is extracted or learned in order to map the input to a predefined emotion model using automatic pattern recognition methods.

and classification or by directly proposing detection regions from input.

One of the first relevant methods for human detection was propose by Viola and Jones [51]. Following a method previously applied to face detection, it employs a cascade structure for efficient detection, and utilizing AdaBoost for automatic feature selection [51].

An important advancement in performance came with the adoption of gradient-based features for describing shape. Dalal and Triggs, popularized the so called histogram of oriented gradient (HOG) features for object detection by showing substantial gains over intensity based features [52]. Since their introduction, the number of variants of HOG features has proliferated greatly with nearly all modern detectors utilizing them in some form [53].

Earlier works on human detection assumed no prior knowledge over the structure of the human body. Arguably one of the most important contributions in this direction was the Deformable Part Models (DPM) [54]. A DPM is a set of parts and connections between the parts which relate to a geometry prior. In the initial proposal by Felzenswalb et al. a discriminative part based approach models unknown part positions as latent variables in a support vector machine (SVM) framework. Local appearance is easier to model than global appearance and training data can be shared across deformations. Some authors argued that there is still no clear evidence for the necessity of components and parts, beyond the case of occlusion handling [55].

In late years a spectacular rise in performance in many pattern recognition problems was brought by training deep neural networks (DNN) with massive amounts of data. According to some authors, the obtained results for human detection are on par with classical approaches like DPM, making the advantage of using such architectures yet unclear [55]. Evenmore, DNN models are known to be very slow, especially when used as sliding-window classifiers which makes it challenging to use for pedestrian detection. One way to alleviate this problem is to use several networks in a cascaded fashion. For example, a smaller, almost shallow network was trained to greatly reduce the initially

large number of candidate regions produced by the sliding window. Then in a second step, only high confidence regions were passed through a deep network obtaining in this way a trade-off between speed and accuracy [56]. The idea of cascading any kind of features of different complexity, including deeply learnt features was addressed by seeking an algorithm for optimal cascade learning under a criterion that penalizes both detection errors and complexity. This made it possible to define quantities such as complexity margins and complexity losses, and account for these in the learning process. This algorithm was shown to select inexpensive features in the early cascade stages, pushing the more expensive ones to the later stages [57]. State-of-the-art accuracy with large gains in speed were reported when detecting pedestrians.

For a comprehensive survey of the human detection literature, the interested reader is referred to [50] and [53], [55].

4.2 Body Pose Detection

Once background has been subtracted, detecting and tracking the human body pose is the second stage in automatic recognition of body gestures of affect. This consists in estimating the parameters of a human body model from a frame or from a sequence of frames, while the position and configuration of the body may change [58].

4.2.1 Body Pose Detection

Due to the high dimensions of the search space and the large number of degrees of freedom, as well as variations of cluttered background, body parameters and illumination, human pose estimation is a challenging task [59], [60]. It also demands avoiding body part penetration and impossible positions.

Body pose estimation can be performed using either model fitting or learning. Model-based methods fit an expected model to the captured data, as an inverse kinematic problem [61], [62]. In this context, the parameters can be estimated based on the tracked feature points, using gradient space similarity matching and the maximum likelihood

resulted from the Markov Chain Monte Carlo approach [63]. However, model-based methods are not robust against local extrema, and require initialization [59].

Performing pose estimation using learning is computationally expensive, because of the high dimensions of the data, and requires a large database of labeled skeletal data. Using poselets for encoding the pose information was proposed in [64], [65]. They utilized SVM for classification of the results of skeletal tracking.

Using parallelism for clustering the appearances was proposed in [66]. Hash-initialized skeletons were tracked through range images using the Iterative Closest Point (ICP) algorithm in [67]. Segmentation and classification of the vertices in a closed 3D mesh into different parts for the purpose of human body tracking was proposed in [68].

Haar-cascade classifiers were trained for the segmentation of head and upper-body parts in [59]. They proposed a hybrid approach which involves model-based fitting as well. The results of the learning and classification procedures were combined with the information from extended distance transform and skin segmentation, in order to fit the data to the skeletal model.

Starting with the DeepPose [69], the field of Human Pose Estimation (HPE) experienced a considerable change from using traditional approaches to developing based on deep networks. In the aforementioned study, the 2D joint coordinates were directly regressed using a deep network. In [39], heatmaps were created based on multiple resolutions of a given image at the same time, in order to obtain the features according to numerous scales.

In fact, utilizing CNNs for 3D HPE [70], [71], [72], [73], [74] is meant to tackle the limitedness of the applicability of classical methods. The latter usually can be trained only based on the few existing 3D pose databases. In [75], the probabilistic 3D pose data were fused by using a multi-stage CNN, which were then considered in order to refine the 2D locations, according to the projected belief maps.

The input and model for a deep learning procedure can be of various types. For example, in [76], simple nearest neighbor upsampling and multiple bottom-up, top-down inferences through stacking numerous hourglasses were proposed, as well as combining a Convolutional Network (ConvNet) with a part-based spatial model. The foregoing approach was computationally efficient as well, which helps take advantage of higher spatial precisions [39].

Dual-source CNN (DS-CNN) [77] is another framework which can be utilized in the context of HPE. Example results obtained by utilizing the foregoing strategy from different studies are shown in Fig. 5.

In [76], each body part was first individually analyzed by the network, which resulted in a heatmap and the corresponding associative embedding tag. After comparing the joint embedding tags, separate pose predictions were made [74]. Heatmaps were taken into account in order to find a 2D bounding box locating the subject using a CNN referred to as 2DPoseNet. Afterward, another CNN, namely, 3DPoseNet, regressed the 3D pose, followed by calculating the global 3D pose and perspective correction based on the camera parameters.

During the last few years, numerous studies have utilized deep learning methods for feature extraction. For example, in [69], [76], [78], seven-layered convolutional DNNs were considered for regressing and representing the joint contexts and locating the body. In [79], high-level global features were obtained from different sources, and then combined through a deep model. In [39], a deep convolutional network was combined with the Markov random field, in order to develop a part-based body detector according to multi-scale features.

In [80], [81], each gesture was modeled based on combinations of large-scale motions of body parts, such as torso, head or limbs, and subtle movements, e.g. those of fingers. The data gathered from the whole time-span were then combined using recursive neural networks (RNN) and long short-term memory (LSTM). Similarly, RNNs with continuously valued hidden layer representations were used for propagating the information over the sequence in studies such as [82], [83], [84].

4.2.2 Tracking the Body Pose

Human actions can differ greatly in their dynamics. According to [85], they can be either periodic (e.g. running, walking or waving) or nonperiodic (e.g. bending), and either stationary (e.g. sitting) or nonstationary/transitional (e.g. skipping as a horizontal motion, and jumping or getting up as vertical motions). In videos we would like to track the human pose along a sequence of consecutive frames. In this way the parameters need to be estimated for every frame, i.e. the position, shape and configuration of the body are found such that they are consistent with the position and configuration of the body at the beginning of the motion capture process [58].

One way to facilitate the tracking of the human body pose is to require subjects to wear special markers, suits or gloves. Therefore, removing the constraints and the requirement of extra hardware were investigated by many researchers, in order to make it possible to perform tracking by using a single camera [58].

After finding the model for the first frame, for each of the next frames, it needs to be readjusted. A common approach to achieve the foregoing goal is to perform iterations such that every time, the model is predicted for the next frame. Expectation Maximization (EM) can be utilized in order to create a human body tracking system, which assigns foreground pixels to the body parts, and then updates their positions according to the data [86], [87]. The parameters of the human body model can be projected onto the optical flow data based on the products of an exponential map [88]. Optical flow and intensity can be used to find human body contours [89]. Then forces have to be applied in order to align the model with the contours extracted from the data. The foregoing process is iterated until the results converge. Configuration spaces of human motions with high dimensions can be handled using a particle filter model [90]. A continuation principle should be utilized based on annealing, in order to introduce the effect of narrow peaks into the fitness function.

An intensely used method for modelling the temporal dimension of the human body is the use of probabilistic directed graphs like Hidden Markov Models (HMM). According to [32], HMMs were used in body gesture recognition [91], [92], automatic sign language interpretation [93]
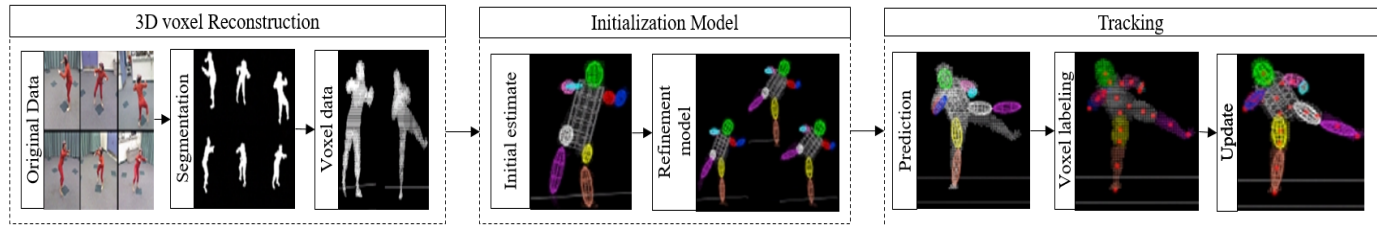
Figure 8. Example of dynamic body pose estimation. From left to right, we depict 3D voxel reconstruction segmentation with the segmentation and extracting the voxel data as input of the Initialitaion model and tracking box. Prediction by using the special filtering, voxel labeling and finally updating by means of the used filtering in the prediction step [58].

and activity analysis [94], [95]. Despite the earlier works, which were based on static pattern matching and dynamic time warping [92], the most recent studies utilize HMMs and dynamic Bayesian networks [96], [97].

Numerous methods of human body tracking may require manual initialization [58]. Otherwise, usually it is necessary to perform a set of calibration movements, in order to help the system to identify the body parts [98], [99]. Other possibilities exist.

4.2.3 3D Body Pose Detection and Tracking

Instead of working on images, voxelized 3D models of human body shapes at each frame can be considered for the purpose of tracking [58]. Voxels provide data, which can be used to create efficient and reliable tracking systems. However, their main disadvantage lies in demanding the additional pre-processing task of converting images to 3D voxel reconstructions, which requires dedicated hardware for a real-time performance [58].

In case of some systems, the input data were obtained using a single camera [100], [101]. Other studies such as [88], [89], [102] proposed to use multiple cameras. A 3D kinematic model of a hand can be tracked based on a layered template representation of self-occlusions [103].

A formulated fully recursive framework for computer vision called DYNA was developed based on 2D blob features from multiple cameras, which were incorporated into a 3D human body model in a probabilistic way [104]. The system provided feedback from the 3D body model to the 2D feature tracking, where the prior probabilities were set using an extended Kalman filter. The foregoing approach can handle behaviors, which are meaningful motions, but can not be explained by passive physics. One example of this kind of tracking is represented in the Fig. 8.

Tapered super-quadrics can be used to create human body models, and track one or more persons in 3D [105]. A constant acceleration kinematic model should be used to predict the position of the body parts in the next frame. The undirected normalized chamfer distances between the image and model contours can be considered to adjust the positions of the head and torso and then those of the arms and legs. Alternative less used pose detection methods are based on silhouette information [106] or motion models [107]. Their reported performances are weak and are not highly flexible, and usually may require additional calculations and processing compared to the rest of the methods we have reviewed. Moreover, they might demand manual contribution from the user. For example, in [108], pedestrian

detection from still images based on silhouette information was proposed. They used low-level gradient data to create, select and learn shapelet features. Although the efficiency of the system was relatively high, later, in studies such as [109], it was shown to result in a less accurate performance than approaches which utilize e.g. HOG features through random forest (RF)-based classification. On the other hand, as the main drawbacks of motion models, they are limited to a certain number of motions, allow only small variations in motions, and cannot handle transitions between different motions [110].

4.3 Representation Learning and Emotion Recognition

The final stage of an EBGR process is building a relevant representation and using it to learn a mapping to the corresponding targets. Depending on the nature of the input, the representation can be static, dynamic or both. Also representation can be geometrical or could include appearance information and can focus on different parts of the body. Moreover, the mapping will then need to be taken into account in order to decide on the most probable class for a given input sample, i.e. to recognize it, which can be performed by using various classification methods. The foregoing topics will be discussed in what follows.

4.3.1 Representation Learning

Gunes et al. [111], [112] detected face and the hands based on the skin color information, and the hand displacement to neutral position was calculated according to the motion of the centroid coordinates. They used the information from the upper body. For example, in a neutral gesture, there is no movement, but in a happy or sad gesture, the body gets extended, and the hands go up, and get closer to the head than normal. More clearly, they defined motion protocols in order to distinguish between the emotions. In the first frame, the body was supposedly in its neutral state, i.e. the hands were held in front of the torso. In the subsequent frames, the in-line rotations of the face and the hands were analyzed. The actions (body modeling) were first coded by two experts. The first and the last frames from each body gesture, which stand for neutral and peak emotional states, respectively, were utilized for training and testing.

Vu et al. [113] considered eight body action units, which represent the movements of the hands, head, legs and waistline. Kipp et al. [114] provided an investigation of a possible correlation between emotions and gestures. The analysis was performed on static frames extracted from

videos representing certain emotional states, as well as emotion dimensions of pleasure, arousal and dominance. The hand shape, palm orientation and motion direction were calculated for all the frames as the features. The magnitudes and directions of the correlations between the expected occurrences and the actual ones were evaluated by finding the correspondences between the dimension pairs, and calculating the resulting deviations.

Glowinski et al. [115] focused on the hands and head as the active elements of an emotional gesture. The features were extracted based on the attack and release parts of the motion cue, which refer to the slope of the line that connects the first value to the first relative extremum and the slope of the line that connects the last value to the last relative extremum, respectively. They also extracted the number of local maxima of the motion cue and the ratio between the maximum and the duration of the largest peak, which were used to estimate the overall impulsiveness of the movement.

Kessous et al. [116] extracted the features from the body and hands. Based on silhouette and hands blobs, they extracted the quantity of motion, silhouette motion images (SMIs) and the contraction index (CI). Velocity, acceleration and fluidity of the hand's barycenter were also computed. Glowinski et al. [117] successfully extended their work using the same database as in [115], where the 3D position, velocity, acceleration and jerk were extracted from every joint of the skeletal structure of the arm. Kipp and Martin [114] used a dimensional method to represent an affect emotional gesture along a number of continuous axes. Three independent bipolar dimensions namely, pleasure, arousal and dominance, were considered in order to define the affective states. The locations of 151 emotional terms were obtained.

In [118], dynamic features were extracted in order to obtain a description of the submotion characteristics, including initial, final and main motion peaks. It was suggested that the timing of the motions greatly represents the properties of emotional expressions. According to [32], these features can be handled based on the concept of motion primitives, i.e. dynamic features can be represented by a number of subactions.

Hirota et al. [119] used the information about the hands, where dynamic time warping (DTW) was utilized to match the time series. Altun et al. [120] considered force sensing resistor (FSR) and accelerometer signals for affect recognition. Lim et al. [121] captured 3D points corresponding to 20 joints at 30 frames per second (fps), where in the recognition stage, 100 previous frames were analyzed in case of every frame.

Saha et al. [122] created skeleton models representing the 3D coordinates of 20 upper body joints, i.e. 11 joints corresponding to the hands, head, shoulders and spine were considered in order to calculate nine features based on the distances, accelerations and angles between them. The distance between the hands and spine, the maximum acceleration of the hands and elbows and the angle between the head, shoulder center and spine were considered as features, making use of static and dynamic information simultaneously.

Camurri et al. [123] utilized five motion cues, namely, QoM, CI, velocity, acceleration and fluidity. Piana et al. [124] proposed 2D and 3D features for dictionary learning. The 3D data were obtained by tracking the subjects, and the 2D data from the segmentation of the images. The spacial data included 3D CI, QoM, motion history gradient (MHG) and barycentric motion index (BMI). Patwardhan et al. [125] utilized 3D static and dynamic geometrical features (skeletal) from the face and upper-body. Castellano et al. [126] considered the velocity and acceleration of the trajectory followed by the hand's barycenter, which was extended in [127], adopting multiple modalities (face, body gesture, speech), and in [115], considering 3D features instead. Vu and et al. [113] used the AMSS [128] in order to find the similarity between the gesture templates and the input samples.

Unfortunately more complex representations are very scarce in emotional body gesture recognition. Chen et al. [129] used HOG on the motion history image (MHI) for finding the direction and speed, and Image-HOG features from bag of words (BOW) to compute appearance features. Another example is the usage of a multichannel CNN for learning a deep representation from the upper part of the body [130]. Finally, Botzheim et al. [131] used spiking neural networks for temporal coding. A pulse-coded neural network approximated the dynamics with the ignition phenomenon of a neuron and the propagation mechanism of the pulse between neurons.

4.3.2 Emotion Recognition

Glowinski et al. [117] showed that meaningful groups of emotions, related to the four quadrants of the valence/arousal space, can be distinguished from representations of trajectories of head and hands from frontal and lateral view of the body. A compact representation was grouped into clusters and used for classifying input into four classes according to the position in the dimensional space namely high-positive(amusement, pride), high-negative (hot-anger, fear, dispair), low-negative (pleasure, relief, interest) or low-negative (cold anger, anxiety, sadness).

Gunes and Piccardi [112] used naive representations from the upper-body to classify body gestures into 6 emotional categories. The categories were groups of the original output space namely: anger-disgust, anger-fear, anger-happiness, fear-sadness-surprise, uncertainty-fear-surprise and uncertainty-surprise. The amount of data and subject diversity were low (156 samples, 3 subjects). A set of standard classifiers were trained and a Bayesian Net provided the best classification results.

Castellano et al. [126] showed comparisons between different classifiers like 1-nearest-neighbor with dynamic time warping (DTW-1NN), J48 decision tree and the Hidden Naive Bayes (HNB) for classifying dynamic representations of body gestures as Anger, Joy, Pleasure or Sadness. The DTW-1NN provided the best results.

Saha et al. [122] identified gestures corresponding to five basic human emotional states, namely, anger, fear, happiness, sadness and relaxation from skeletal geometrical features. They compared binary decision tree (BDT), ensemble tree (ET), k-nearest neighbour (KNN) and SVM, obtaining the best results by using ET.

Many studies modeled parts of the body independently for action analysis [32]. For example, in [132], actions based on arms, head and torso were modeled independently. On the contrary, a structural body model was proposed in [133]. They defined a tree-based description of the body parts, where each activity corresponded to a node, based on the parts engaged in performing it.

Context information such as background were introduced by Kosti et al. [134]. They used a two low-rank filter CNN for extracting features from both body and background and fusing them for recognizing 26 emotions and intensity values of valence, arousal and dominance. Some examples of the emotions they targeted are peace, affection, fatigue and pain.

Although body gestures are important part of human communication, often they are a supplement of other reflexive behavior forms such as facial expression speech, or context. Studies in applied psychology showed that human recognition of facial expressions is influenced by the body expression and by the context [135]. Integration of verbal and nonverbal communication channels creates a system in which the message is easier to understand. Expanding the focus to several expression forms can facilitate research on emotion recognition as well as human-machine interaction.

In literature, there are just a few examples concerning speech and gestures recognition. In [136] the authors focused on uncovering the emotional effect on the interrelation between speech and body gestures. They used prosody and mel-frequency cepstral coefficient (MFCC) [137], [138] for speech with three types of body gestures: head motion, lower and upper body motions to study the relationship between the two communication channels affected by the emotional state. Additionally, they proposed a framework for modeling the dynamics of speech-gesture interaction.

In [113] the authors also presented a bi-modal approach (gestures and speech) for recognition of four emotional states: happiness, sadness, disappointment, and neutral. Gestures recognition module fused two sources: video and 3D acceleration sensors. Speech recognition module was based on Julius [139] - open source software. The outputs from speech and gestures based recognition were fused by using weight criterion and best probability plus majority vote fusion methods. Fifty Japanese words (or phrases) and 8 types of gestures recorded from five participants were used to validate the system. Performance of the classifier indicated better results for bi-modal than each of the uni-modal recognition system.

Regarding fusing gestures and faces the literature is also scarce. Gunes et al. [140] proposed a bi-model emotion recognition method from body and face. Features were extracted from two streams of video of the face and the body and fusions were performed at feature and decision level improving results with respect to individual feature space. A somewhat similar approach was proposed by Caridakis et al. [141] combining face, body and speech modalities and early and late-fusion followed by simple statistical classification approaches with considerable improvement in results.

Psaltis et al. [142] introduced a multi-modal late fusion structure that could be used for stacked generalization on noisy databases. Surprise, hapiness, anger, sadness and fear were recognized from facial action units and high representation of the body gestures. The multi-modal approach provided better recognition than results from each of the mono-modal.

A very interesting study on multi-modal automatic emotion recognition was presented in [116]. The authors constructed a database consisting of audio-video recordings of people interacting with an agent in a specific scenario. Ten people of different gender, using several different native languages including French, German, Greek and Italian pronounced a sentence in 8 different emotional states. Facial expression, gesture and acoustic features were used with an automatic system based on a Bayesian classifier. Results obtained from each modality were compared with the fusion of all modalities. Combining features into multi-modal sets resulted in a large increase in the recognition rates, by more than 10% when compared to the most successful unimodal system. Furthermore, the authors proved that the best results were obtained for gesture and speech feature merger.

4.4 Applications

Applications of automatic recognition of gesture based expression of affect are mainly of three types [143], [144], [145]. The first type consists of systems that detect the emotions of the users. The second type includes actual or virtual animated conversational agents, such as robots and avatars. They are expected to act similarly to humans when they are supposed to have a certain feeling. The third type includes systems that really feel the emotions. For example, these systems have applications in video telephony [146], video conferencing and stress-monitoring tool, violence detection [147], [148], [149], video surveillance [150], and animation or synthesis of life-like agents [148] and automatic psychological research tools [150]. All the three types have been extensively discussed in the literature. However, this paper concentrates on affect detection only.

Automatic multi-modal emotion recognition systems can utilize sources of information that are based on face, voice and body gesture, at the same time. Thus they can constitute an important element of perceptual user interfaces, which may be utilized in order to improve the ease of use of online shops. They can also have applications in pervasive perceptual man-machine interfaces, which are used in intelligent affective machines and computers that understand and react to human emotions [151]. If the system is capable of combining the emotional and social aspects of the situations for making a decision based on the available cues, it can be a useful assistant for humans [152].

5 DATA

We further present main public databases of gesture based expressions of affect useful for training EGBR systems. We discuss RGB, Depth and bi-modal of RGB + Depth databases in Sec. 5.1, 5.2 and 5.3, respectively. The reader is referred to Table 2 for an overview of the main characteristics of the databases and to Fig. 9 for a selection of database samples.

Figure 9. Selected samples from databases containing gesture based expressions of affect: (a) FABO [21], (b) GEMEP-FERA [115], [153], [154], (c) Theater [114], (d) HUMAINE [116], [126], [127], [155], (e) LIRIS-ACCEDE [156], [157], (f) MSR-Action 3D [158].

5.1 RGB

One of the first body language databases with affect annotations was made publicly available by Gunes and Piccardi [112]. The database contains 206 samples with six basic emotions, as well as four more states, namely, neutral, anxiety, boredom and uncertainty. 156 samples were used for training, and 50 samples for the test.

Castellano et al. [126], [127] collected affective body language data consisting of 240 gestures [159]. Their database is a part of the HUMAINE database [155]. There were six male and four female participants, i.e. 10 in total. They acted eight emotions, i.e. anger, despair, interest, pleasure, sadness, irritation, joy and pride, equally distributed in the valence arousal space. However, they focused on four emotions, i.e. anger, joy, pleasure and sadness. A camera filmed the full body of the subjects from the front view at a rate of 25 fps. In order to accelerate the silhouette extraction, they used a uniform dark background.

The Geneva Multi-modal Emotion Portrayals (GEMEP) database [115] contains more than 7000 audio-video portrayals of emotional expressions. It includes 18 emotions portrayed by 10 actors. 150 portrayals were systematically chosen based on ratings by experts and non-experts, which resulted in the best recognition of the emotional intentions. Their analysis was on the basis of 40 portrayals selected from the mentioned set. They were chosen such that they represent four emotions, namely, anger, joy, relief and sadness. Each of these emotions is from one quadrant of the two main affective dimensions, i.e. arousal and valence.

The Theater corpus was introduced by Kipp and Martin [114], based on two movie versions of the play *Death of a Salesman*, namely, DS-1 and DS-2.

Vu et al. [113] considered eight types of gestures that are present in the home party scenario of the mascot robot system. The database involved five participants, i.e. four males and one female. Their ages ranged from 22 to 30 years. They were from three different nationalities: Japanese, Chinese, and Vietnamese.

A subset of the LIRIS-ACCEDE video database [157] was created in [156], which contains upper bodies of 64 subjects, including 32 males and 32 females, with six basic emotions. Their ages were between 18 and 35 years.

5.2 Depth

The GEMEP-FERA database, which was introduced by Baltrušaitis et al. [153], is a subset of the GEMEP corpus. The training database was created by 10 actors. In the test database, six actors participated. From these actors, three were common with the training database, but the other three were new. The database consists of short videos of the upper body of the actors. The average length of the videos is 2.67 seconds. The videos do not start with a neutral state.

The database created by Saha et al. [122] involved 10 subjects. The age of the subjects ranged from 20 to 30. The subjects were stimulated by five different emotions, namely, anger, fear, happiness, sadness, and relaxation. These emotions caused the subjects to take different gestures accordingly. Each subject was filmed at a frame rate of 30 fps, for 60 seconds. Next, the Cartesian coordinates of the body joints were processed.

In [125], six basic emotions, namely, anger, surprise, disgust, sad, happy and fear, were acted by 15 subjects. The subjects were between 25 to 45 years old. Five subjects were female, and the rest were male. In addition, five subjects were Americans, and the rest were Asians. The lighting conditions were controlled, and the poses of the bodies of the subjects were completely frontal. The subjects' distances from the camera were from to 1.5 to 4 meters.

The UCFKinect [160] was collected using Kinect and skeleton estimation from [161]. 16 subjects, including 13 males and three females, participated in the recordings. All the subjects were between 20 and 35 years old. Each of them performed 16 actions such as balance, punch or run and repeated it five times. In total, 1280 actions were recorded. The 3D coordinates of 15 joints were calculated for each frame. The data on the background and the clothes were not included in the calculations, and only the data on the skeleton was extracted.

The MSR Action 3D consists of twenty actions, namely, high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up and throw.

5.3 Bi-modal: RGB + Depth

The database created by Psaltis et al. [142] contains facial expressions that frequently appear in games. The subjects acted five basic emotions, namely, anger, fear, happiness, sadness and surprise. They considered a neutral emotion class for labeling the samples that do not present any motion, and cannot be classified under any of the basic emotions. 15 subjects participated to create 450 videos. Each video starts with an almost neutral state, and evolves toward a peak emotional state. The labels were assigned based on the emotion that the subject was asked to perform, not on the actually performed movements. The whole duration of every video is 3 seconds. Before acting each emotion, a video presenting the required emotion was shown to the subjects, and then the subjects performed the movements with their own styles five times. The database is divided into three parts. One part only contains facial expressions, another part only contains body gestures, and the last part contains both face and body data. They used a dense-ASM tracking algorithm for tracking the features and extracting the AUs. In order to evaluate the performance of the proposed method, they applied it to the FERA database as well.

The emoFBVP database [162] includes multi-modal recordings of actors, i.e. face, body gesture, voice and physiological signals. Audiovisual information of three different expressions, i.e. intensities, of 23 emotions are included, as well as tracking of facial features and skeletal tracking. 10 professional actors participated in acquiring the data. Each recording was repeated six times. Three recordings were in a standing position, and the others in a seated position. To date, this database offers the most diverse range of emotional body gestures in the literature, but right now it is not available. Finally, the specifications of all available databases are summarized in Table 2. The list of emotions that have been considered in each of the databases is provided in Table 3. A sample image from each of the databases can be seen in Fig. 9.

6 DISCUSSION

In this section we discuss the different aspects of automatic emotional body gesture recognition presented in this work. We start with the collections of data currently available for the community. Then, the discussion mainly focuses on representation building and emotion recognition from gestures. This includes the categories of mostly used features, taking advantage of complementarity by using multi-modal approaches and most common pattern recognition methods and target spaces.

6.1 Data

Majority of freely accessible data sets contain acted expressions. This type of material is usually composed of high quality recordings, with clear undistorted emotion expression. The easiness of acquiring such recordings opens a possibility of obtaining several samples from a single person. The conventional approach to collect acted body language databases is to let actors present a scene portraying particular emotional states. Professionals are able to immerse in an emotion they perform, which may be difficult for ordinary people. This kind of samples are free from uncontrollable influences and usually they do not require additional evaluation and labeling processes. However, some researchers emphasize that these types of recordings may lead to creating a set of many redundant samples, as there is a high dependency on actors skills and his or her ability to act out the same emotional state differently. Another argument against such recordings states that they do not reflect real world conditions. Moreover, acted emotions usually comprise of basic emotions only, whereas in real life emotions are often weak, blurred, occur as combinations, mixtures, or compounds of primary emotions. Wherefore current trends indicate that spontaneous emotions are preferable for research. There is another method to record emotional body movements in natural situations. One can use movies, TV programs such as talk shows, reality shows or live coverage. This type of material might not always be of satisfactory quality (background noise, artifacts, overlapping, etc.) and may obscure the exact nature of recorded emotions. Moreover, collections of spontaneous samples must be evaluated by human decision makers or professional behaviorists to determine the gathered emotional states. Nonetheless it does not guarantee objective, genuinely independent assessments. Additionally, copyright reasons might make it difficult to use or disclose movies or TV recordings. An accurate solution for sample acquisition may be provoking an emotional reaction using staged situations, which has been already used in emotion recognition from speech or mimics. Appropriate states may be induced using imaging methods (videos, images), stories or computer games. This type of recordings are preferred by psychologists, although the method can not provide desirable effects as reaction to the same stimuli may differ. Similarly to spontaneous speech recordings, triggered emotional samples should be subjected to a process of labeling. Ethical or legal reasons often prohibit to use or make them publicly available. Taking into account above mentioned issues, real-life emotion databases are rarely available to the public, and a good way of creating and labelling such samples is still open to question.

The process of choosing appropriate representation of emotional states is intricate. It is still debatable how detailed and which states should be covered. Analyzing Table 3 one can observe how broad affective spectrum has been used in various types of research. Most authors focus on sets containing six basic emotions (according to Ekman's model). Sadness and anger occur in majority of databases. Fear, surprise and disgust are also commonly used. However, there are quite a lot of affective states that are not consistently represented in the available databases. Some examples (see Tab. 3) are uncertainty, unconcern, aghastenss, shame, tenderness, etc.

There is a lack of consistency in the taxonomy used for naming the affective states. For example both joy and happiness, are used interchangeably depending on the database. It is difficult to evaluate whether these are the same or different states. Joy is more beneficial, as it is less transitory than happiness and is not tied to external circumstances. Therefore, it is possible that there is a misunderstanding in naming: while happiness may be caused by down to earth experiences, material objects, joy needs rather spiri-

Table 2
Main characteristics of a selected list of publicly available databases for recognizing gesture based expression of affect.

Reference	Name	Device	Body parts	Modality	#Emotions	#Gestures	#Subjects	#Females	#Males	#Sequences	#Samples	FR[1] (fps)	Background	AVL[2] (s)
Gunes et al., 2006 [21]	FABO	Digital camera	Face and body	Visual	10	NA	23	12	11	23	206	15	Uniform blue	∼3600
Glowinski et al., 2008 [115]	GEMEP	Digital camera	Face and body	Audiovisual	18	NA	10	5	5	1260	>7000	25	Uniform dark	NA
Castellano et al., 2007 [126]	HUMAINE	Camera	Face and body	Audiovisual	8	8	10	4	6	240	240	25	Uniform dark	NA
Gavrilescu, 2015 [156]	LIRIS-ACCEDE	Camera	Face and upper body	Visual	6	6	64	32	32	NA	NA	NA	Nonuniform	60
Baltrusaitis et al., 2011 [153]	GEMEP-FERA	Kinect	Upper body	Visual	5	7	10	NA	NA	289	NA	30	Uniform dark	2.67
Fothergill et al., 2012 [163]	MSRC-12	Kinect	Whole body	Depth	NA	12	30	40%	60%	594	6244	30	Uniform white	40
Masood et al., 2011 [160]	UCFKinect	Kinect	Whole body	Depth	NA	16	16	NA	NA	NA	1280	30	NA	NA
Li et al., 2010 [158]	MSR-Action 3D	Structured light	Whole body	Depth	NA	20	7	NA	NA	NA	567	15	Nonuniform	NA

Table 3
Labels included in a selected list of the databases. F = FABO [21], G = GEMEP [115], T = T heater [114], H = HUMAINE [126], LA = LIRIS-ACCEDE [156], GF = GEMEP-FERA [153].

Database	F	G	T	H	LA	GF	Frequency
Sadness	●	●	●	●	●	●	6
Anger	●		●	●	●	●	5
Anxiety	●	●	●				3
Disgust	●	●			●		3
Fear	●				●	●	3
Surprise	●	●					3
Boredom	●		●				2
Happiness	●					●	2
Interest		●		●			2
Contempt		●					2
Despair		●	●				2
Irritation		●		●			2
Joy			●			●	2
Pleasure		●		●			2
Relief		●				●	2
Admiration		●	●				1
Neutral		●					1
Pride		●		●			1
Shame		●					1
Aghastness		●					1
Amazement			●				1
Amusement		●					1
Boldness				●			1
Comfort				●			1
Dependency			●				1
Disdain			●				1
Distress			●				1
Docility			●				1
Elation		●					1
Excitement			●				1
Exuberance			●				1
Fatigue	●						1
Gratefulness			●				1
Hostility			●				1
Indifference			●				1
Insecurity			●				1
Nastiness			●				1
Panic Fear		●					1
Rage		●					1
Relaxation			●				1
Respectfulness			●				1
Satisfaction			●				1
Tenderness		●					1
Uncertainty	●						1
Unconcern			●				1

Table 4
Summary of a few multi-modal emotion recognition methods.
S=Speech, F=Face, H=Hands, B=Body.

Reference	Modalities	#samples	#emotions	Representation
Gunes Piccardi [112]	F + B	206	6	Motion protocols
Castellano's et al. [126]	B	240	4	Multi cue
Castellano et al. [127]	B	240	4	Multi cues
Glowinski et al. [115]	B	40	4	Multi cues
Kipp Martin [114]	B	#119	6	PAD
Kessous et al. [116]	S + F + B	NA	8	Multi cues
Vu et al. [113]	S + B	5	4	Motion protocols
Gavrilescu [156]	B + H	384	6	Motion protocols

tual experiences, gratitude, and thankfulness, thus may be difficult to evoke and act. These misunderstandings may be also a result of translations. Such issues will reoccur until a consistent taxonomy of emotions will be presented, so far there is no agreement among experts even on the very definition of primary states. Moreover, due to the heterogeneity of described databases, comparison of their quality is problematic. With just several public accessible emotional databases and with the addition of the above described issues, comparison of detection algorithms becomes a challenging task. There is clearly space and necessity of creation of more unified emotional state databases.

6.2 Representation Learning and Emotion Recognition

Representation Learning. The large majority of the methods developed to recognize emotion from body gestures use geometrical representations. A great part of these methods build simple static or dynamic features related to the co-ordinates of either joints of kinematic models or of parts of the body like head, hands or torso. Some of the most used features are displacements [112], orientation of hands [114] motion cues like velocity and acceleration [115], [117], [118], [120], [123], [126], shape information and silhoutte [116], smoothness and fluidity, periodicity, spatial extent and kinetic energy, among others. While most descriptors are very simple there are also examples of slightly more advanced descriptors like Quantity of Motion (QoM measures of the amount of motion in a sequence), Silhoutte Motion Images (SMI contains information about the changes of the shape and position of the silhouette), Contraction Index (CI measures the level of contraction or expansion of the body), and Angular Metrics for Shape Similarity (AMSS) [113], [124].

Considering dynamic features such as acceleration, movement gain and velocity, or at least combining them with static features, usually leads to higher recognition
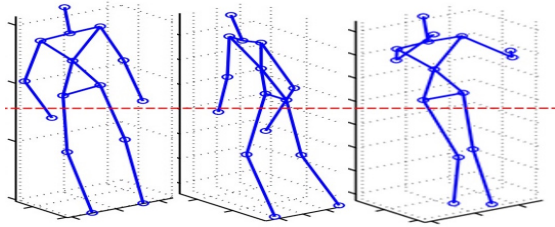
Figure 10. Sample upper and lower body postures [164].

Table 5
Comparison of the effect of using various classification methods and different numbers of emotion classes based on HUMAINE EU-IST database [127].

Classifier	Performance (%)	#classes
1NN-DTW	53.70	4
J48 or Quinlan's C4.5	56.48	4
Hidden Naive Bayes	51.85	4

rates than relying solely on the latter, since they result in a richer representation and since some emotional traits are expressed mostly in the dynamics of the human body.

Most of the methods proposed focus on the upper body (head, neck, shoulders, arms, hands and hand fingers) [114], [119], [122], hands [115], arm [117], body and hands [116], full body [113], [124]. Upper body and lower body parts are represented in Fig. 10.

Among all different parts of the body, in the context of body gesture recognition, numerous studies have focused on hand gestures, which requires hand segmentation and tracking. The features that can be extracted from the hands include palm orientation, hand shape, elbow, wrist, palms and shoulder joints, hand shape and motion direction, which are analyzed independently from the body, in order to calculate the motion of the hand and the individual fingers along each of the axes. The motions of the hand are measured with the body as the reference. The motions of the body itself are found in terms of the changes of the pose of the upper body, i.e. its inclinations to the left, right, forward or backward.

Complex learnt representations for recognizing emotion from body gestures are very scarce, mostly because there is a lack of big volumes of labelled data (see Tab. 2) for learning such representations in a supervised way. Two of the very few works that uses deep learning representations for body emotion recognition are multichannel CNN from upper body [130] and spiking neural networks for temporal coding [131]. As previously discussed in Sec. 6.1 there is a lack of consistent taxonomy for the output spaces in the various databases published to date. This results in considerable fragmentation of the data and makes transfer learning techniques difficult. Even though not explored yet in the literature, unsupervised learning might be interesting for pre-training general representations of the moving human body, before tuning to more specific emotion oriented models.

Emotion Recognition. There is a tendency in the literature to reduce the output spaces for simplifying the recognition problem. This has been done either by grouping emotions into quadrants of a dimensional emotion space [117] or by grouping emotions based on similarity of their apperance [112]. In general, most methods have focused on recognizing basic emotions like Anger, Joy, Pleasure, Sadness and Fear [122], [126]. Though not dominant, methods that target richer output spaces also exist [134].

Another popular approach is to show extensive comparison between sets of standard classifiers like decision trees, k-NNs and SVMs. The results of using numerous classifiers and different numbers of emotion classes based

on two different databases are summarized in Table 5 and 6, respectively.

According to Table 5, J48 [165] has the best performance between three used classifiers tested on HUMAINE database. This work used just four labels of the mentioned database. According to Table 6 the best performance on a different database, with 10 subjects and 5 labels, is achieved by ensemble tree classification methods. Moreover, the different methods with their performances are represented as a chart for emotional gesture recognition in Fig. 11.

More complete representations of the body can also be used in a more meaningful way. Particularly interesting are structural models where different parts of the body are indepedtly represented and contribute to a final decision over the emotion which takes into account predefined priors [133]. Going even further, additional information from the context, like the background could be used as well to refine final decision [134].

Different body language components (gestures, faces) together with speech carry affective information and complementary processing have obvious advantages. A consistent part of the literature uses multiple representations of the body in a complementary way to recognize emotion. For example, there are works that combine body with speech [113], [136] and with face [140], [141]. Regardless of the fusion techniques used, all these methods report improvements of results backing the hypothesis that there is considerable complementarity in different modalities and its exploration is fruitfull. Also, it has already been previously commented that a more complete body representation is also helpful in this respect (for example, upper and lower body considered together). Unfortunately research in multimodal emotion recognition remains rather scarce and simplistic. The few works that exists mostly focus in simplistic fusion techniques from shallow representations of body and face or body and speech. Even though all methods report important improvements over monomodal equivalents, this potential remains largely unexplored. The reader is refered to Table 4 for a selected set of studies that have used body representations together with representations of other modalities for recognizing emotion.

The number of emotion classes affects the performance of a given classifier as well. Usually, reducing the number of classes from a given database should increase the performance. The best recognition rate, i.e. 93%, is obtained by considering five emotion classes and using neural networks. It should be noted that low-quality samples or features may degrade the performance, and cause a violation of the expected trend.

Approaches to train and test emotional gesture recog-

Table 6
Comparison of the effect of using various classification methods and different numbers of emotion classes based on the recorded samples by Kinect. The database included by 10 subjects in the age group of 25±5 years [122].

Classifier	Performance (%)	#classes
Ensemble tree	90.83	5
Binary decision tree,	76.63	5
K-NN	86.77	5
SVM	87.74	5
Neural network	89.26	5

Figure 11. Performances (%) of different emotion recognition methods based on the different databases.

nition systems are investigated based on the existing literature, where a certain portion of the database is used for training, and the rest is left for testing. Some of the proposed techniques present superior performances on specific databases, i.e. they have led to accuracy rates higher than 90%. However, in order to ensure that the system is reliable, it needs to be tested against different types of data, including various conditions of background, e.g. dark, light, uniform and nonuniform. Moreover, it is worth paying attention that different training and testing strategies may result in different performance rates.

7 CONCLUSION

In this paper we defined a general pipeline of Emotion Body Gesture Recognition methods and detailed its main blocks. We have briefly introduced important pre-processing concepts like person detection and body pose estimation and detailed a large variety of methods that recognize emotion from body gestures grouped along important concepts such as representations learning and emotion recognition methods. For introducing the topic and broadening its scope and implications we defined emotional body gestures as a component of body language, an essential type of human social behavior. The difficulty and challenges of detecting general patterns of affective body language are underlined. Body language varies with gender and has important cultural dependence vital issues for any researcher willing to publish data or methods in this field.

In general the representations used remain shallow. Most of them are naive geometrical representations, either skeletal or based on independently detected parts of the body.

Features like motion cues, distances, orientations or shape descriptors abound. Even though recently we can see deep meaningful representations being learned for facial analysis for affect recognition a similar approach for a more broader affective expression of humans is still to be developed in the case of body analysis. For sure the scarcity of body gesture and multimedia affective data is playing a very important role, problem that recently is starting to be overcome in the case of facial analysis. An additional problem is that while in the case of facial affective computing there has been a quite clear consensus of the output space (primitive facial expressions, facial Action Units and recently more comprehensive output spaces) in the case of general affective expressions in a broader sense such consensus does not exist. A proof in this sense is the variety of labels proposed in the multitude of publicly available data, some of them following redundant or confusing taxonomies.

In general, for comprehensive affective human analysis from body language, emotional body gesture recognition should learn from emotional facial recognition and clearly agree on sufficiently simple and well defined output spaces based on which to publish large high quality amounts of labelled and unlabelled data that could serve for learning rich deep statistical representations of the way the affective body language looks like.

REFERENCES

[1] A. Pease, B. Pease, The definitive book of body language. Peace International, Peace International, 2004.

[2] C. Darwin, P. Prodger, The expression of the emotions in man and animals, Oxford University Press, USA, 1998.

[3] P. Ekman, F. Wallace, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologist Press, 1978.

[4] N. Smrtić, Asertivna komunikacija i komunikacija u timu, Ph.D. thesis, Polytechnic of Međimurje in Čakovec. Management of tourism and sport. (2015).

[5] K. Brow, Kinesics. Encyclopedia of Language and Linguistics 2nd Edition, Elsevier Science, 2005.

[6] B. Pease, A. Pease, The definitive book of body language, Bantam, 2004.

[7] D. Rosenstein, H. Oster, Differential facial responses to four basic tastes in newborns, Child development (1988) 1555–1568.

[8] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, et al., Universals and cultural differences in the judgments of facial expressions of emotion., Journal of personality and social psychology 53 (4) (1987) 712.

[9] B. d. Gelder, Why Bodies? Twelve Reasons for Including Bodily Expressions in Affective Neuroscience., hilosophical Transactions of the Royal Society B: Biological Sciences 364 (364) (2009) 3475–3484.

[10] A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: A survey, TAC 4 (1) (2013) 15–33.

[11] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, D. Kulić, Body movements for affective expression: A survey of automatic recognition and generation, TAC 4 (4) (2013) 341–359.

[12] C. A. Corneanu, M. O. Simon, J. F. Cohn, S. E. Guerrero, Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications, IEEE transactions on pattern analysis and machine intelligence 38 (8) (2016) 1548–1568.

[13] S. Escalera, V. Athitsos, I. Guyon, Challenges in multi-modal gesture recognition, in: Gesture Recognition, Springer, 2017, pp. 1–60.

[14] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, S. Escalera, Deep learning for action and gesture recognition in image sequences: A survey, in: Gesture Recognition, Springer, 2017, pp. 539–578.

[15] J. F. Iaccino, Left brain-right brain differences: Inquiries, evidence, and new approaches, Psychology Press, 2014.

[16] H. Ruthrof, The body in language, Bloomsbury Publishing, 2015.

[17] P. Molchanov, S. Gupta, K. Kim, J. Kautz, Hand gesture recognition with 3d convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 1–7.

[18] A. Pease, B. Pease, The Definitive Book of Body Language: how to read others' attitudes by their gestures, Hachette UK, 2016.

[19] A. W. Siegman, S. Feldstein, Nonverbal behavior and communication, Psychology Press, 2014.

[20] H. Gunes, M. Piccardi, Fusing face and body gesture for machine recognition of emotions, in: Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on, IEEE, 2005, pp. 306–311.

[21] H. Gunes, M. Piccardi, A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior, in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, Vol. 1, IEEE, 2006, pp. 1148–1153.

[22] H. Gunes, C. Shan, S. Chen, Y. Tian, Bodily expression for automatic affect recognition, Emotion recognition: A pattern analysis approach (2015) 343–377.

[23] D. Efron, Gesture and environment.

[24] A. Kendon, The study of gesture: Some remarks on its history, in: Semiotics 1981, Springer, 1983, pp. 153–164.

[25] Dimension of body language , http://westsidetoastmasters.com/resources/book_of_body_language/toc.html/, [Online; accessed 19-June-2017].

[26] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (3-4) (1992) 169–200.

[27] L. A. Camras, H. Oster, J. J. Campos, K. Miyake, D. Bradshaw, Japanese and american infants' responses to arm restraint., Developmental Psychology 28 (4) (1992) 578.

[28] https://www.udemy.com/body-language-basics-in-business-world, accessed: 2017-12-15.

[29] C. Frank, Stand Out and Succeed: Discover Your Passion, Accelerate Your Career and Become Recession-Proof, Nero, 2015.

[30] R. W. Simon, L. E. Nath, Gender and emotion in the united states: Do men and women differ in self-reports of feelings and expressive behavior?, American journal of sociology 109 (5) (2004) 1137–1176.

[31] U. Hess, S. Senécal, G. Kirouac, P. Herrera, P. Philippot, R. E. Kleck, Emotional expressivity in men and women: Stereotypes and self-perceptions, Cognition & Emotion 14 (5) (2000) 609–642.

[32] D. Bernhardt, Emotion inference from human body motion, Ph.D. thesis, University of Cambridge (2010).

[33] T.-y. Wu, C.-c. Lian, J. Y.-j. Hsu, Joint recognition of multiple concurrent activities using factorial conditional random fields, in: Proc. 22nd Conf. on Artificial Intelligence (AAAI-2007), 2007.

[34] M. A. Fischler, R. A. Elschlager, The representation and matching of pictorial structures, IEEE Transactions on computers 100 (1) (1973) 67–92.

[35] T. Tung, T. Matsuyama, Human motion tracking using a color-based particle filter driven by optical flow, in: The 1st International Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA'08, 2008.

[36] P. F. Felzenszwalb, D. McAllester, Object detection grammars., in: ICCV Workshops, 2011, p. 691.

[37] W. Zhang, J. Shen, G. Liu, Y. Yu, A latent clothing attribute approach for human pose estimation, in: Asian Conference on Computer Vision, Springer, 2014, pp. 146–161.

[38] A. Newell, Z. Huang, J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, in: Advances in Neural Information Processing Systems, 2017, pp. 2274–2284.

[39] J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: Advances in neural information processing systems, 2014, pp. 1799–1807.

[40] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[41] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, Human pose estimation with iterative error feedback, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4733–4742.

[42] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, M. R. Wróbel, Modeling emotions for affect-aware applications, Cover and title page designed by ESENCJA Sp. z oo (2015) 55.

[43] P. Ekman, Universal and cultural differences in facial expression of emotion, Nebr. Sym. Motiv. 19 (1971) 207–283.

[44] J. Russell, A. Mehrabian, Evidence for a three-factor theory of emotions, J. Research in Personality 11 (1977) 273–294.

[45] R. Plutchik, The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, American scientist 89 (4) (2001) 344–350.

[46] P. Ekman, Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique, Psychol. Bull. 115 (2) (1994) 268–287.

[47] M. Greenwald, E. Cook, P. Lang, Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli, J. Psychophysiology (3) (1989) 51–64.

[48] D. Watson, L. A. Clark, A. Tellegen, Development and validation of brief measures of positive and negative affect: The PANAS scales, JPSP 54 (1988) 1063–1070.

[49] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, TPAMI 31 (1) (2009) 39–58.

[50] D. T. Nguyen, W. Li, P. O. Ogunbona, Human detection from images and videos: a survey, Pattern Recognition 51 (2016) 148–175.

[51] P. Viola, M. J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, in: null, IEEE, 2003, p. 734.

[52] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: European conference on computer vision, Springer, 2006, pp. 428–441.

[53] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, TPAMI 34 (4) (2012) 743–761.

[54] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: CVPR, IEEE, 2008, pp. 1–8.

[55] R. Benenson, M. Omran, J. Hosang, B. Schiele, Ten years of pedestrian detection, what have we learned?, arXiv preprint arXiv:1411.4304.

[56] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. S. Ogale, D. Ferguson, Real-time pedestrian detection with deep network cascades., in: BMVC, 2015, pp. 32–1.

[57] Z. Cai, M. Saberian, N. Vasconcelos, Learning complexity-aware cascades for deep pedestrian detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3361–3369.

[58] I. Mikić, M. Trivedi, E. Hunter, P. Cosman, Human body model acquisition and tracking using voxel data, International Journal of Computer Vision 53 (3) (2003) 199–223.

[59] A. Kar, Skeletal tracking using microsoft kinect, Methodology 1 (2010) 1–11.

[60] G. Anbarjafari, S. Izadpanahi, H. Demirel, Video resolution enhancement by using discrete and stationary wavelet transforms with illumination compensation, Signal, Image and Video Processing 9 (1) (2015) 87–92.

[61] C. Barron, I. A. Kakadiaris, Estimating anthropometry and pose from a single image, in: CVPR, Vol. 1, IEEE, 2000, pp. 669–676.

[62] C. J. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, in: CVPR, Vol. 1, IEEE, 2000, pp. 677–684.

[63] M. Siddiqui, G. Medioni, Human pose estimation from a single view point, real-time range sensor, in: CVPRW, IEEE, 2010, pp. 1–8.

[64] L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1365–1372.

[65] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, H.-P. Seidel, Motion capture using joint skeleton tracking and surface estimation, in: CVPR, IEEE, 2009, pp. 1746–1753.

[66] D. Ramanan, D. A. Forsyth, Finding and tracking people from the bottom up, in: CVPR, Vol. 2, IEEE, 2003, pp. II–II.

[67] D. Grest, J. Woetzel, R. Koch, Nonlinear body pose estimation from depth images, in: DAGM-Symposium, Vol. 5, Springer, 2005, pp. 285–292.

[68] E. Kalogerakis, A. Hertzmann, K. Singh, Learning 3d mesh segmentation and labeling, in: ACM Transactions on Graphics (TOG), Vol. 29, ACM, 2010, p. 102.

[69] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.

[70] M. Lin, L. Lin, X. Liang, K. Wang, H. Cheng, Recurrent 3d pose sequence machines, arXiv preprint arXiv:1707.09695.

[71] H. Coskun, Human pose estimation with cnns and lstms, Master's thesis, Universitat Politècnica de Catalunya (2016).

[72] S. Li, A. B. Chan, 3d human pose estimation from monocular images with deep convolutional neural network, in: Asian Conference on Computer Vision, Springer, 2014, pp. 332–347.

[73] C.-H. Chen, D. Ramanan, 3d human pose estimation= 2d pose estimation+ matching, arXiv preprint arXiv:1612.06524.

[74] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3d human pose estimation in the wild using improved cnn supervision, 3DV 1 (2) (2017) 5.

[75] D. Tome, C. Russell, L. Agapito, Lifting from the deep: Convolutional 3d pose estimation from a single image, arXiv preprint arXiv:1701.00295.

[76] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, Springer, 2016, pp. 483–499.

[77] X. Fan, K. Zheng, Y. Lin, S. Wang, Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1347–1355.

[78] W. Yang, W. Ouyang, H. Li, X. Wang, End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3073–3082.

[79] W. Ouyang, X. Chu, X. Wang, Multi-source deep learning for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2329–2336.

[80] N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. Taylor, F. Nebout, A multi-scale approach to gesture detection and recognition, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 484–491.

[81] A. Metallinou, A. Katsamanis, S. Narayanan, Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information, Image and Vision Computing 31 (2) (2013) 137–152.

[82] J. Wan, S. Escalera, X. Baro, H. J. Escalante, I. Guyon, M. Madadi, J. Allik, J. Gorbova, G. Anbarjafari, Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges, in: ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV, Vol. 4, 2017.

[83] P. R. Khorrami, How deep learning can help emotion recognition, Ph.D. thesis, University of Illinois at Urbana-Champaign (2017).

[84] D. Wu, N. Sharma, M. Blumenstein, Recent advances in video-based human action recognition using deep learning: A review, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE, 2017, pp. 2865–2872.

[85] L. Wang, D. Suter, Learning and matching of dynamic shape manifolds for human action recognition, IEEE Transactions on Image Processing 16 (6) (2007) 1646–1661.

[86] E. Hunter, Visual estimation of articulated motion using the expectation-constrained maximization algorithm, University of California, San Diego, 1999.

[87] E. Hunter, P. Kelly, R. Jain, Estimation of articulated motion using kinematically constrained mixture densities, in: Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE, IEEE, 1997, pp. 10–17.

[88] C. Bregler, J. Malik, Tracking people with twists and exponential maps, in: CVPR, IEEE, 1998, pp. 8–15.

[89] Q. Delamarre, O. Faugeras, 3d articulated models and multiview tracking with physical forces, Computer Vision and Image Understanding 81 (3) (2001) 328–357.

[90] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, in: CVPR, Vol. 2, IEEE, 2000, pp. 126–133.

[91] A. D. Wilson, A. F. Bobick, Parametric hidden markov models for gesture recognition, TPAMI 21 (9) (1999) 884–900.

[92] Y. Wu, T. S. Huang, Vision-based gesture recognition: A review, in: Gesture Workshop, Vol. 1739, Springer, 1999, pp. 103–115.

[93] T. Starner, A. Pentland, Real-time american sign language recognition from video using hidden markov models, in: Motion-Based Recognition, Springer, 1997, pp. 227–243.

[94] T. Mori, Y. Segawa, M. Shimosaka, T. Sato, Hierarchical recognition of daily human actions based on continuous hidden markov models, in: FG, IEEE, 2004, pp. 779–784.

[95] N. Oliver, E. Horvitz, A. Garg, Layered representations for human activity recognition, in: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, IEEE Computer Society, 2002, p. 3.

[96] Y. Du, F. Chen, W. Xu, Y. Li, Recognizing interaction activities using dynamic bayesian network, in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, Vol. 1, IEEE, 2006, pp. 618–621.

[97] R. El Kaliouby, P. Robinson, Generalization of a vision-based computational model of mind-reading, Affective computing and intelligent interaction (2005) 582–589.

[98] I. A. Kakadiaris, D. Metaxas, Three-dimensional human body model acquisition from multiple views, International Journal of Computer Vision 30 (3) (1998) 191–218.

[99] G. K. Cheung, T. Kanade, J.-Y. Bouguet, M. Holler, A real time system for robust 3d voxel reconstruction of human motions, in: CVPR, Vol. 2, IEEE, 2000, pp. 714–720.

[100] C. Sminchisescu, B. Triggs, Covariance scaled sampling for monocular 3d body tracking, in: CVPR, Vol. 1, IEEE, 2001, pp. I–I.

[101] S. Ioffe, D. Forsyth, Human tracking with mixtures of trees, in: ICCV, Vol. 1, IEEE, 2001, pp. 690–695.

[102] I. A. Kakadiaris, D. Metaxas, Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection, in: CVPR, IEEE, 1996, pp. 81–87.

[103] J. M. Rehg, T. Kanade, Model-based tracking of self-occluding articulated objects, in: Computer Vision, 1995. Proceedings., Fifth International Conference on, IEEE, 1995, pp. 612–617.

[104] C. R. Wren, Understanding expressive action, Ph.D. thesis, Massachusetts Institute of Technology (2000).

[105] D. M. Gavrila, L. S. Davis, 3-d model-based tracking of humans in action: a multi-view approach, in: CVPR, IEEE, 1996, pp. 73–80.

[106] A. Senior, Real-time articulated human body tracking using silhouette information, in: Proc. of IEEE Workshop on Visual Surveillance/PETS, 2003, pp. 30–37.

[107] R. Urtasun, P. Fua, 3d human body tracking using deterministic temporal motion models, in: European conference on computer vision, Springer, 2004, pp. 92–106.

[108] P. Sabzmeydani, G. Mori, Detecting pedestrians by learning shapelet features, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.

[109] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, P. H. Torr, Randomized trees for human pose detection, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[110] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer vision and image understanding 104 (2) (2006) 90–126.

[111] H. Gunes, M. Piccardi, T. Jan, Face and body gesture recognition for a vision-based multimodal analyzer, in: Proceedings of the Pan-Sydney area workshop on Visual information processing, Australian Computer Society, Inc., 2004, pp. 19–28.

[112] H. Gunes, M. Piccardi, Affect recognition from face and body: early fusion vs. late fusion, in: 2005 IEEE international conference

on systems, man and cybernetics, Vol. 4, IEEE, 2005, pp. 3437–3443.

[113] H. A. Vu, Y. Yamazaki, F. Dong, K. Hirota, Emotion recognition based on human gesture and speech information using rt middleware, in: Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, IEEE, 2011, pp. 787–791.

[114] M. Kipp, J.-C. Martin, Gesture and emotion: Can basic gestural form features discriminate emotions?, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009, pp. 1–8.

[115] D. Glowinski, A. Camurri, G. Volpe, N. Dael, K. Scherer, Technique for automatic emotion recognition by body gesture analysis, in: CVPRW, IEEE, 2008, pp. 1–6.

[116] L. Kessous, G. Castellano, G. Caridakis, Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis, Journal on Multimodal User Interfaces 3 (1-2) (2010) 33–48.

[117] D. Glowinski, M. Mortillaro, K. Scherer, N. Dael, G. V. A. Camurri, Towards a minimal representation of affective gestures, in: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, 2015, pp. 498–504.

[118] G. Castellano, Movement expressivity analysis in affective computers: from recognition to expression of emotion, Unpublished doctoral dissertation). Department of Communication, Computer and System Sciences, University of Genoa, Italy.

[119] K. Hirota, H. A. Vu, P. Q. Le, C. Fatichah, Z. Liu, Y. Tang, M. L. Tangel, Z. Mu, B. Sun, F. Yan, et al., Multimodal gesture recognition based on choquet integral, in: Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, IEEE, 2011, pp. 772–776.

[120] K. Altun, K. E. MacLean, Recognizing affect in human touch of a robot, Pattern Recognition Letters 66 (2015) 31–40.

[121] A. Lim, H. G. Okuno, The mei robot: towards using motherese to develop multimodal emotional intelligence, IEEE Transactions on Autonomous Mental Development 6 (2) (2014) 126–138.

[122] S. Saha, S. Datta, A. Konar, R. Janarthanan, A study on emotion recognition from body gestures using kinect sensor, in: Communications and Signal Processing (ICCSP), 2014 International Conference on, IEEE, 2014, pp. 056–060.

[123] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci, G. Volpe, Toward real-time multimodal processing: Eyesweb 4.0, in: Proc. Artificial Intelligence and the Simulation of Behaviour (AISB) 2004 Convention: Motion, Emotion and Cognition, Citeseer, 2004, pp. 22–26.

[124] S. Piana, A. Stagliano, A. Camurri, F. Odone, A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition, in: IDGEI International Workshop, 2013.

[125] A. Patwardhan, G. Knapp, Augmenting supervised emotion recognition with rule-based decision model, arXiv preprint arXiv:1607.02660.

[126] G. Castellano, S. D. Villalba, A. Camurri, Recognising human emotions from body movement and gesture dynamics, in: International Conference on Affective Computing and Intelligent Interaction, Springer, 2007, pp. 71–82.

[127] G. Castellano, L. Kessous, G. Caridakis, Emotion recognition through multiple modalities: face, body gesture, speech, in: Affect and emotion in human-computer interaction, Springer, 2008, pp. 92–103.

[128] T. Nakamura, K. Taki, H. Nomiya, K. Uehara, Amss: A similarity measure for time series data, IEICE Transactions on Information and Systems 91 (2008) 2579–2588.

[129] S. Chen, Y. Tian, Q. Liu, D. N. Metaxas, Recognizing expressions from face and body gesture by temporal normalized motion and appearance features, Image and Vision Computing 31 (2) (2013) 175–185.

[130] P. Barros, D. Jirak, C. Weber, S. Wermter, Multimodal emotional state recognition using sequence-dependent deep hierarchical features, Neural Networks 72 (2015) 140–151.

[131] J. Botzheim, J. Woo, N. T. N. Wi, N. Kubota, T. Yamaguchi, Gestural and facial communication with smart phone based robot partner using emotional model, in: World Automation Congress (WAC), 2014, IEEE, 2014, pp. 644–649.

[132] F. Lv, R. Nevatia, Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, ECCV (2006) 359–372.

[133] S. Vacek, S. Knoop, R. Dillmann, Classifying human activities in household environments, in: Workshop at the International Joint Conference on Artificial Intelligence (IJCAI), 2005.

[134] R. Kosti, J. M. Alvarez, A. Recasens, A. Lapedriza, Emotion recognition in context, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[135] J. Van den Stock, R. Righart, B. De Gelder, Body expressions influence recognition of emotions in the face and voice., Emotion 7 (3) (2007) 487.

[136] Z. Yang, S. S. Narayanan, Analysis of emotional effect on speech-body gesture interplay, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[137] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, G. Anbarjafari, Fusion of classifier predictions for audio-visual emotion recognition, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 61–66.

[138] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, G. Anbarjafari, Audio-visual emotion recognition in video clips, IEEE Transactions on Affective Computing.

[139] Open-Source Large Vocabulary CSR Engine Julius , http://http://julius.osdn.jp/en_index.php/, [Copyright 2014 Julius development team; accessed 7-July-2017].

[140] H. Gunes, M. Piccardi, Bi-modal emotion recognition from expressive face and body gestures, Journal of Network and Computer Applications 30 (4) (2007) 1334–1345.

[141] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, K. Karpouzis, Multimodal emotion recognition from expressive faces, body gestures and speech, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2007, pp. 375–388.

[142] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. C. Apostolakis, K. Dimitropoulos, P. Daras, Multimodal affective state recognition in serious games applications, in: Imaging Systems and Techniques (IST), 2016 IEEE International Conference on, IEEE, 2016, pp. 435–439.

[143] R. W. Picard, R. Picard, Affective computing, Vol. 252, MIT Press, 1997.

[144] R. W. Picard, Affective computing for hci., in: HCI (1), 1999, pp. 829–833.

[145] R. W. Picard, Affective computing: from laughter to ieee, TAC 1 (1) (2010) 11–17.

[146] J. Cassell, A framework for gesture generation and interpretation, Computer vision in human-machine interaction (1998) 191–215.

[147] M. Pantic, L. J. M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, TPAMI 22 (12) (2000) 1424–1445.

[148] M. Pantic, L. J. Rothkrantz, Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE 91 (9) (2003) 1370–1390.

[149] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, R. Sukthankar, Violence detection in video using computer vision techniques, in: Computer Analysis of Images and Patterns, Springer, 2011, pp. 332–339.

[150] A. Pentland, Looking at people: Sensing for ubiquitous and wearable computing, TPAMI 22 (1) (2000) 107–119.

[151] R. W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, TPAMI 23 (10) (2001) 1175–1191.

[152] B. Reeves, C. Nass, How people treat computers, television, and new media like real people and places, CSLI Publications and Cambridge university press Cambridge, UK, 1996.

[153] T. Baltrušaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson, R. Picard, Real-time inference of mental states from facial expressions and upper body gestures, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 909–914.

[154] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, K. Scherer, Meta-analysis of the first facial expression recognition challenge, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42 (4) (2012) 966–979.

[155] E. Douglas-Cowie, C. Cox, J.-C. Martin, L. Devillers, R. Cowie, I. Sneddon, M. McRorie, C. Pelachaud, C. Peters, O. Lowry, et al., The humaine database, in: Emotion-Oriented Systems, Springer, 2011, pp. 243–284.

[156] M. Gavrilescu, Recognizing emotions from videos by studying facial expressions, body postures and hand gestures, in: Telecom-

munications Forum Telfor (TELFOR), 2015 23rd, IEEE, 2015, pp. 720–723.

[157] Y. Baveye, E. Dellandrea, C. Chamaret, L. Chen, Liris-accede: A video database for affective content analysis, TAC 6 (1) (2015) 43–55.

[158] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: CVPRW, 2010, pp. 9–14.

[159] I. Humaine, Human-machine interaction network on emotion, 2004-2007 (2008).

[160] S. Z. Masood, C. Ellis, A. Nagaraja, M. F. Tappen, J. J. LaViola, R. Sukthankar, Measuring and reducing observational latency when recognizing actions, in: ICCVW, 2011, pp. 422–429.

[161] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Communications of the ACM 56 (1) (2013) 116–124.

[162] H. Ranganathan, S. Chakraborty, S. Panchanathan, Multimodal emotion recognition using deep learning architectures, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–9.

[163] S. Fothergill, H. Mentis, P. Kohli, S. Nowozin, Instructing people for training gestural interactive systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 1737–1746.

[164] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, R. Sukthankar, Exploring the trade-off between accuracy and observational latency in action recognition, International Journal of Computer Vision 101 (3) (2013) 420–436.

[165] H. Chauhan, A. Chauhan, Implementation of decision tree algorithm c4. 5, International Journal of Scientific and Research Publications 3 (10).

Ciprian Adrian Corneanu got his BSc in Telecommunication Engineering from Télécom SudParis in 2011 and his MSc in Computer Vision from Universitat Autónoma de Barcelona in 2015. Currently he is a Ph.D. student at the Universitat de Barcelona and a fellow of the Computer Vision Center, UAB. His main research interests include face and behavior analysis, affective computing, social signal processing and human computer interaction.

Tomasz Sapiński received his M.Sc. degree in Computer Science from Faculty of Technical Physics, Information Technology and Applied Mathematics at Łodz University of Technology. Currently he is Ph.D. student at Institute of Mechatronics and Information Systems, Łodz University of Technology. His main research topics are: multi-modal emotion recognition and practical applications of virtual reality.

Sergio Escalera obtained the P.h.D. degree on Multi-class visual categorization systems at Computer Vision Center, UAB. He obtained the 2008 best Thesis award on Computer Science at Universitat Autònoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at UB, CVC, and the Barcelona Graduate School of Mathematics. He is an associate professor at the Department of Mathematics and Informatics, Universitat de Barcelona. He is an adjunct professor at Universitat Oberta de Catalunya, Aalborg University, and Dalhousie University. He has been visiting professor at TU Delft and Aalborg Universities. He is also a member of the Computer Vision Center at UAB. He is series editor of The Springer Series on Challenges in Machine Learning. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events. His research interests include, between others, statistical pattern recognition, affective computing, and human pose recovery and behavior understanding, including multi-modal data analysis.

Fatemeh Noroozi received her B.Sc. in Computer Engineering, Software, from Shiraz University, Iran. Her thesis was entitled "Modeling of Virtual Organizations Integrating on Physical Core based on a Service-oriented Architecture". Afterwards, she received her M.Sc. in Mechatronics Engineering from the University of Tehran, Iran. Her thesis was entitled "Developing a Real-time Virtual Environment for Connecting to a Touching Interface in Dental Applications ". Currently, she is a PhD student at the University of Tartu, Estonia, working on "multi-modal Emotion Recognition based Human-robot Interaction Enhancement".
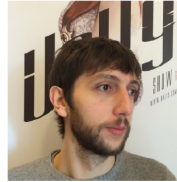

Dorota Kamińska graduated in Automatic Control and Robotics and completed postgraduate studies in Biomedical image processing and analysis at Łodz University of Technology. She received her PhD degree from Faculty of Electrical, Electronic, Computer and Control Engineering at Łodz University of Technology in 2014. The topic of her thesis was "Emotion recognition from spontaneous speech". She gained experience during the TOP 500 Innovators programme at Haas School of Business, University of California in Berkeley. Currently she is an educator and scientist at Institute of Mechatronics and Information Systems. She is passionate about biomedical signals processing for practical appliances. As a participant of many interdisciplinary and international projects, she is constantly looking for new challenges and possibilities of self-development.

Gholamreza Anbarjafari heads the intelligent computer vision (iCV) research lab in the Institute of Technology at the University of Tartu. He is an IEEE Senior member and the Vice Chair of the Signal Processing / Circuits and Systems / Solid-State Circuits Joint Societies Chapter of the IEEE Estonian section. He received the Estonian Research Council Grant (PUT638) and the Scientific and Technological Research Council of Turkey (Proje 1001 - 116E097) in 2015 and 2017, respectively. He has been involved in many international industrial projects. He is expert in computer vision, human-robot interaction, graphical models and artificial intelligence. He is an associated editor of several journals such as SIVP and JIVP and have been lead guest editor of several special issues on human behaviour analysis. He has supervised over 10 MSc students and 7 PhD students. He has published over 100 scientific works. He has been in the organizing committee and technical committee of conferences such as ICOSST, ICGIP, SIU, SampTA, FG and ICPR. He is organizing a challenge and a workshop on in FG17, CVPR17, and ICCV17.